

A. Background on de Bruijn Sequences and Linear-Feedback Shift Registers

Here, we give a brief introduction to the generation of de Bruijn sequences by linear-feedback shift registers¹ (LFSRs). A more thorough mathematical treatment will be described in a separate paper (Philippakis *et al.*, manuscript in preparation). As stated in the main text, a de Bruijn sequence is a circular string of length 4^k that contains every k -mer exactly once when overlaps are considered. It can be proved that de Bruijn sequences exist for all values of k and all possible alphabet sizes and, in fact, that there are a large number of such sequences for any choice of k and alphabet size². In the current application we utilized a class of de Bruijn sequences generated by LFSRs. Such sequences are known to have provable pseudo-randomness properties¹ that are advantageous, since they guarantee that any trends observed in the data are not a result of how the sequences were generated.

Consider the Boolean field $Z_2 = \{0,1\}$, and its associated addition (+) and multiplication (\times) operators. In order to generate a de Bruijn sequence of order k over the DNA alphabet {A,C,G,T}, we first recursively generate a sequence of length $2^{2k}-1$ over the Boolean field. Here, the i 'th element of the sequence $S = (s_1 s_2 s_3 \dots)$ is generated from the preceding $2k$ elements by the equation:

$$s_i = a_1 s_{i-1} + a_2 s_{i-2} + \dots + a_{2k} s_{i-2k}$$

If the coefficients $\{a_1, a_2, \dots, a_{2k}\}$ are chosen so that the corresponding polynomial:

$$a_1 x^1 + a_2 x^2 + \dots + a_{2k} x^{2k}$$

is primitive³ over Z_2 , then the sequence S generated by this recursive equation will have periodicity $2^{2k}-1$ and will contain every subsequence of length $2k$ over the Boolean alphabet except the sub-sequence that contains $2k$ 0's¹.

In order to transform S into a deBruijn sequence over the DNA alphabet, take the following embedding, Π , into $Z_2 \times Z_2$:

$$\Pi(A) = (0,0)$$

$$\Pi(C) = (0,1)$$

$$\Pi(G) = (1,0)$$

$$\Pi(T) = (1,1)$$

If one then takes the deBruijn sequence of order $2k$ over the Boolean field and transforms pairs of letters with this embedding into a sequence over the DNA alphabet while transliterating both reading frames (Philippakis *et al.*, manuscript in preparation), then the resulting sequence over the DNA alphabet will contain all variants of length k except the sequence of k A's, (this element can be added by inserting an extra "A" to one of the subsequences containing $k-1$ A's).

We have further developed the theory of LFSRs in order to judiciously choose primitive polynomials such that the generated de Bruijn sequences, in addition to representing all contiguous k -mers, also represent various gapped patterns of k -mers. Such de Bruijn sequences have the dual advantage of ensuring that binding sites for transcription factors (TFs) with gapped motifs are well-covered by the array and also ensuring that k' -mers (where $k' > k$) are regularly sampled, facilitating interpolation to k' -mers not represented on the array (Philippakis *et al.*, manuscript in preparation). In the present application, we

have utilized primitive polynomials that represented all 10-mers with a single nucleotide gap.

B. Enrichment Score and Motif Construction

Consider a TF whose binding sites are of width k . For two distinct k -mers a and b , we desire to utilize the measured signal intensities for features on our microarray to determine the relative preferences of the TF for a and b . As outlined in the main text, we have identified five sources of variability, described below, that confound the direct assignment of these preferences by inspecting any pair of features (say, F_a and F_b) containing matches to a and b . 1) The relative *efficiencies* with which features F_a and F_b are double-stranded. 2) The geometric *location* of features F_a and F_b on the microarray, as there are occasional non-uniformities in the protein binding or labeling reactions. 3) The *position* of the binding sites a and b within F_a and F_b , as sites positioned more proximally to the glass surface tend to have lower signal intensities (see **Supplementary Fig. 4**). 4) The *orientation* of binding sites a and b (*i.e.*, forward orientation or reverse complement), as we have observed that for some TFs there is a preferred orientation (see **Supplementary Fig. 4**). 5) The *flanking sequence* within which a and b are embedded, as the presence of additional moderate or low affinity sites besides those under consideration can increase the observed signal intensity; similarly, if the true width of the binding site is wider than k , then the positions immediately flanking the binding sites could positively or negatively impact the observed signal intensity.

Our experimental design and data analysis approaches were specifically designed to alleviate these five confounding variables. For (1), we have performed primer extensions with labeled nucleotides (Cy3-conjugated dUTP), allowing normalization by the amount of double-stranded DNA on a given feature (described in **Methods**). For (2), we have developed a method of local-averaging in order to lessen these location effects (also described in **Methods**). Effects (3)-(5) were somewhat more challenging to remove. The approach that we have initially adopted (and that is utilized in this manuscript), is two-fold: first, to exploit the fact that k -mers (for $k < 10$) are represented multiple times on the array and second, to perform a replicate experiment on a second microarray whose features were constructed using a different de Bruijn sequence (the location, position, and flanking sequence of each candidate binding site will therefore be different on this second array, further allowing these effects to be lessened by averaging over many replicates with different positions, locations, and flanking sequences). This approach, however, requires the development of a statistical metric that can determine the relative preferences of the TF for binding sites a and b from the ensemble of all features containing a and b on both replicate arrays.

In developing this metric, we deemed non-parametric statistics based on the ordering of signal intensities to be most appropriate, as the distribution of signal intensities is unknown (it is somewhat log-normal, but has a heavy right tail) and may contain outliers due to the confounding factors mentioned above. Additionally, we wanted this metric to be invariant of sample size, so that binding sites occurring on different numbers of features could be compared on the same scale. Previously, we utilized a rank-based

statistic that was a modified form of the Wilcoxon-Mann-Whitney (WMW) scaled to be independent of sample size which we called “area,” as it geometrically corresponded to the difference between the areas of the foreground and background detection rate curves⁴ (here “foreground” features are those containing a match to a given candidate binding site and “background” features are all other features). Mathematically, this area is expressed as:

$$area = \frac{1}{B + F} \left[\frac{\rho_B}{B} - \frac{\rho_F}{F} \right] \quad (\text{Eqn. 1})$$

where B and F are the sizes of the background and foreground, respectively, and ρ_B and ρ_F are the sums of the background and foreground ranks. Unfortunately, this metric is sensitive to low outliers, as a few features with high ranks could greatly increase the value of ρ_F . This is especially problematic in the present application, since the bottom half of features have very similar signal intensities but very different ranks, potentially exacerbating the effect of a few low outliers. To address this, we modified this statistic by only considering the top half of features in the foreground and background that have the highest signal intensities. The foreground and background features were then pooled and the area statistic was computed using **Eqn. 1**; we henceforth refer to this modified area statistic as the *enrichment score*. We note that such modifications of the rank-sum statistic have been studied under the generalized setting of L-statistics, along with their asymptotic limiting distributions^{5, 6}.

Enrichment scores for all contiguous 8-mers and 9-mers (after identifying reverse complements) are published on our website, http://the_brain.bwh.harvard.edu. In addition to these values, however, we sought to develop a more compact representation of the DNA binding specificities of TFs from the measured signal intensities in the form of a position weight matrix. Our approach to motif construction consists of four distinct steps which are described below; note that key elements of this approach are that all features and their rankings are used (as opposed to simply running a motif-finding program such as AlignACE⁷ on probes whose features were above an arbitrary threshold), and that motifs are built by combining data from two replicate arrays utilizing distinct de Bruijn sequences.

First, recall that our de Bruijn sequences were constructed to represent not only all contiguous 10-mers exactly once, but also all 10-mers with a gap size of 1 (e.g., 10-mers of the form AnCAGATTACG, ATnAGATTACG, ATCnGATTACG, ... , ATCAGATTAnG). Thus, all 8-mers with up to three gaps are represented 16 times on each de Bruijn sequence (32 times after identifying reverse complements for non-palindromic 8-mers), and so have a sufficiently high copy-number to allow averaging over some of the aforementioned confounding effects. We began by computing enrichment scores for all 8-mers containing up to 3 gapped positions. Note that these enrichment scores were first computed for each potentially gapped 8-mer separately for each array. These enrichment scores were then averaged over the two arrays to identify the 8-mer with the highest enrichment score. This 8-mer is henceforth referred to as the “seed” of the motif. The seeds of the five motifs used in this study were:

Cbf1: GTCACGTG

Zif268: GCGTGGGC

Ceh-22: CCACTTGA
 Oct1: TATGCAAA
 Rap1: GGTGTnnGGG

Second, for each of the 8 positions constituting the seed of the motif, we inspected each of the four nucleotide variants at that position in order to quantify the relative preferences of the TF for each of these variants. Let $F_{i,j,p}$ be the set of all features on array $i \in \{1,2\}$ that contain a match to the variant of the seed that has letter $j \in \{A,C,G,T\}$ at position $p \in \{1,2,\dots,8\}$, or that has a match to the reverse complement of this variant; thus, using Cbf1 as an example, $F_{1,A,3}$ denotes that set of all features on array 1 that match GTAACGTG or CACGTTAC. Next, let $\overline{F_{i,j,p}}$ be the collection of all features on array i that do not match the variant of the seed that has letter j at position p , but that do contain a match to one of the other three binding site variants at that position or its reverse complement; again using Cbf1 as an example, $\overline{F_{1,A,3}}$ is the set of all features that match one of {GTCACGTG, GTGACGTG, GTTACGTG, CACGTGAC, CACGTCAC, CACGTAAC}. Finally, let $S_{i,j,p}$ and $\overline{S_{i,j,p}}$ be the signal intensities for features in $F_{i,j,p}$ and $\overline{F_{i,j,p}}$, and let $N_{i,j,p}$ and $\overline{N_{i,j,p}}$ be the number of features in $F_{i,j,p}$ and $\overline{F_{i,j,p}}$. For each array individually, and for each value of $p \in \{1,2,\dots,8\}$, we determine the relative importance of variant j by pooling features in $F_{i,j,p}$ and $\overline{F_{i,j,p}}$, sorting them by their values $S_{i,j,p}$ and $\overline{S_{i,j,p}}$, and then computing:

$$\psi_{i,j,p} = \frac{1}{N_{i,j,p} + \overline{N_{i,j,p}}} \left[\frac{\rho_{i,j,p}}{N_{i,j,p}} - \frac{\overline{\rho_{i,j,p}}}{\overline{N_{i,j,p}}} \right] \quad (\text{Eqn. 2})$$

where $\rho_{i,j,p}$ is the rank-sum of features in $F_{i,j,p}$ and $\overline{\rho_{i,j,p}}$ is the rank-sum of features in $\overline{F_{i,j,p}}$ (note that in the computation of $\psi_{i,j,p}$ in **Eqn. 2**, we did not drop the lower half of data points in $F_{i,j,p}$, as the background here is much smaller than that used to find the optimal seed, thereby simultaneously lessening the effects of low outliers making the computed values more sensitive to sample variance). Finally, values of $\psi_{i,j,p}$ were averaged over the two arrays in order to generate a single measure of the relative preference of the letter j at position p of the motif, giving $\psi_{j,p} = (\psi_{1,j,p} + \psi_{2,j,p})/2$.

Third, we sought to extend our motif beyond the eight positions in the seed of the motif. For this, we first transformed the values $\psi_{j,k}$ into probabilities by utilizing a Boltzmann distribution^{8,9}:

$$P(j, p) = \frac{\exp(\gamma * \psi_{j,p})}{\sum_{j' \in \{A,C,G,T\}} \exp(\gamma * \psi_{j',p})} \quad (\text{Eqn. 3})$$

Since the values of $\psi_{j,p}$ are in the range $\psi_{j,p} \in [-0.5, 0.5]$, we chose $\gamma = \ln(10,000)$ since, if a given letter is maximally enriched among the brightest features (i.e., $\psi_{j,p} = 0.5$) and the other three variants are equally depleted (i.e., $\psi_{j',p} = -0.167$), then $P(j,p) \approx 0.99$. We note that this choice of γ only affects the visual representation of the motif, but does not affect

the ordering of relative preferences induced on k -mers. In order to determine the least informative position within the 8-mer seed, we used these derived probabilities to compute the relative entropy of each position⁹:

$$H(p) = \sum_{j \in \{A,C,G,T\}} P(j,p) \log_2 \left(\frac{P(j,p)}{0.25} \right) \quad (\text{Eqn. 4})$$

and selected the position p with minimum $H(p)$.

Next, we extended the motif beyond the 8 positions considered in its seed. Consider the 7-mer corresponding to the original 8-mer seed after dropping the least informative position, and consider all additional positions that, when added to this 7-mer, yield an 8-mer with no more than three gapped positions. For example, if the original 7-mer seed had the pattern of informative positions 11011111 where 1's represent fixed positions in the seed and 0's represent positions that are ignored, then the extensions of this 7-mer that are considered are {1001101111, 1011011111, 1110111111, 1101111111, 1101111101, 11011111001}. At each of these additional positions we repeat the computations performed in Step 2 by inspecting all four variants and calculating $\psi_{i,j,p}$ at these positions. Here also, the values of $\psi_{i,j,p}$ at these extended positions are computed separately for each array, and then averaged to give $\psi_{j,p}$.

Fourth, all values of $\psi_{j,p}$ at the 8 positions within the original seed and also at all extended positions were transformed into probabilities $P(j,p)$ using **Eqn. 3**. We then used the program enoLOGOS¹⁰ to display them graphically.

We stress that although this is a first-generation approach, it has the following two desirable features: 1) it utilizes information on the ranks of all features in constructing the motif, instead of arbitrarily choosing a subset and weighting them equally to construct the motif; and 2) it is able to systematically combine measurements made from arrays constructed using different de Bruijn sequences; moreover, it can be expanded to incorporate an arbitrarily large numbers of such distinct arrays, which is likely to be useful as the technology is further developed. Although the basic analysis presented here can undoubtedly be expanded and improved (for example, by building motifs from multiple initial seeds which are then merged), we anticipate that an incorporation of these two considerations is likely to be a key component of future methods of motif construction from data generated by such compact, universal microarrays.

Finally, we note that in **Figure 3c**, we demonstrated interdependencies in the Cbf1 motif by fixing the sequence “CACGTG” and then varying the first two positions of “nnCACGTG”. There, we computed the WMW statistic (i.e., the enrichment score) for each of the 16 variants, taking as a background the remaining 15 variants; thus, this generalizes the method that we used for single positions to the case of dinucleotides. As above, this was done for each array individually, and the resulting values were averaged for the two arrays.

C. Normalization of PBM Signal Intensity by Cy3 dUTP Signal Intensity

Primer extension was performed using unlabeled dATP, dCTP, dGTP, dTTP, and a small amount of Cy3-conjugated dUTP, as described in **Methods**. Consequently, unless there were large differences in the yield of double-stranding from spot to spot, the Cy3 signal was expected to be linearly proportional to the number of adenines in the template strand. The Cy3 signal was indeed roughly proportional to the number of adenines, yet a strong dependence on the local sequence context of adenines was observed. For example, an adenine following a pyrimidine resulted in substantially higher signal intensity than an adenine following a purine.

To determine whether the context-dependent Cy3 variation was due to true sequence-dependent differences in double-stranding efficiency or to an incorporation bias against modified nucleotides, we designed a set of twenty control sequences for our microarray. Each contains the primer sequence, followed by a variable 16-nt sequence, followed by the Zif268 binding site embedded in constant flanking sequence. For the variable region, we used 20 different sequences that had the same mononucleotide frequencies (i.e., four each of the letters A,C,G,T) but differing dinucleotide frequencies. Each sequence was present at 16 identical replicate spots on the microarray. Primer extension reactions and Zif268 PBM experiments were carried out as described in the text. These 20 different sequences gave reproducibly different Cy3 signal intensities after primer extension ($P = 2.9 \times 10^{-126}$; ANOVA), yet their Zif268 PBM signal intensities were quite consistent ($P = 0.86$; ANOVA). We reasoned that the consistency of the PBM signal intensities indicates that the yield of double-stranding is independent of the dinucleotide composition of the template, and consequently that the incorporation of modified Cy3 dUTP, but not unlabeled dTTP, is dependent on dinucleotide composition. Furthermore, we tested 10 different second-order de Bruijn sequences (containing 1 of each dinucleotide but with different trinucleotide combinations) in a similar fashion. These 10 sequences also gave reproducibly different Cy3 signal intensities ($P = 2.6 \times 10^{-25}$; ANOVA), but similar Zif268 PBM signal intensities ($P = 0.841$; ANOVA), suggesting that the incorporation of modified Cy3 dUTP is also dependent on the trinucleotide composition of the template.

The Cy3 intensities of the 40,330 variable spots containing the de Bruijn sequence were used to compute regression coefficients for the relative contributions of all trinucleotide combinations to the total signal. Regressing over trinucleotides gave a substantially better approximation than over dinucleotides, although the addition of a fourth position contributed negligibly (**Supplementary Fig. 9**). Using these regression coefficients, a ratio of observed-to-expected Cy3 signal intensity was calculated for each sequence. The PBM signal intensity of each spot was divided by this ratio, and all spots with observed-to-expected Cy3 signal intensity <0.5 were removed from further consideration.

D. Determining Protein Concentration

GST-tagged TFs were purified and concentrated as described in **Methods**. The molarities of all purified proteins were determined by Western blot using a dilution series of recombinant GST (Sigma). Equal volumes of sample and known concentrations of GST were suspended in 1x NuPAGE LDS Sample Buffer (Invitrogen), heated to 95°C for 5 minutes, and loaded on a NuPAGE 4-12% Bis-Tris gel (Invitrogen). Samples were

electrophoresed at 200 V for 25 minutes and then transferred to nitrocellulose membranes at 30 V for 3 hours according to the manufacturer's protocols. Membranes were then labeled and developed using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Pierce) according to the manufacturer's protocols. Primary antibody (anti-GST produced in rabbit; Sigma) was added at 20 ng/ml, and secondary antibody (horseradish-peroxidase-conjugated anti-rabbit IgG produced in goat; Pierce) was added at 5 ng/ml. Film was scanned and analyzed using Quantity One version 4.5.0 software (Bio-Rad) by interpolating the sample concentrations from the GST standard curve.

E. Surface Plasmon Resonance

The following oligonucleotides (Integrated DNA Technologies) were used in the Biacore experiments described in the text to ascertain the equilibrium binding constants of Cbf1 for several variant binding sites:

GTCACGTG:

5'-nnnnnnnnnnnnnggtcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

ATCACGTG:

5'-nnnnnnnnnnnnngatcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

TTCACGTG:

5'-nnnnnnnnnnnnngttcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

GGCACGTG:

5'-nnnnnnnnnnnnngggcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

AGCACGTG:

5'-nnnnnnnnnnnnngagcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

TGCACGTG:

5'-nnnnnnnnnnnnngtcacgtgggnnnnnnnnnnnnnnnngaaaggatgggtgacgacgcg-3'

Control Flow Cell:

5'-nngaaaggatgggtgacgacgcg-3'

Because the last six positions of the binding site are palindromic, Cbf1 can potentially bind in either orientation. These probes were specifically designed so that the binding site is followed by GG, as CC appeared in PBMs to be one of the most disfavored dinucleotides to occupy the first two positions of the binding site. This ensured that Cbf1 would most often bind in the orientation containing the binding site variant of interest.

Prior to their use Biacore experiments, these oligonucleotides were converted to double-stranded DNA by primer extension using the following biotinylated, HPLC-purified primer (Integrated DNA Technologies):

5'-cgcgtcgacccatccttc-3'.

References:

1. Golomb, S. Shift Register Sequences. (Aegean Park Press, Laguna Hills, CA; 1967).
2. Gross, J.L. & Yellen, J. Handbook of Graph Theory. (CRC Press, New York; 2004).
3. Stewart, I. Galois Theory. (Chapman & Hall, London, UK; 1989).
4. Philippakis, A.A. et al. Expression-guided *in silico* evaluation of candidate *cis* regulatory codes for *Drosophila* muscle founder cells. *PLoS Computational Biology* **in press** (2006).
5. Bjerve, S. Error Bounds for Linear Combinations of Order Statistics. *Annals of Statistics* **5**, 357-369 (1977).
6. Gastwirth, J.L. Percentile Modifications of Two Sample Rank Tests. *Journal of the American Statistical Association* **60**, 1127-1141 (1965).
7. Hughes, J.D., Estep, P.W., Tavazoie, S. & Church, G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* **296**, 1205-1214 (2000).
8. Berg, O.G. & von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* **193**, 723-750 (1987).
9. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
10. Workman, C.T. et al. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* **33**, W389-392 (2005).